Toward Promoting Prosocial Interactions between Humans with Autonomous Agents

XINYUE HU, University of California, Irvine, Irvine, California, USA
SHASHANK MEHROTRA, TERUHISA MISU, and KUMAR AKASH, Honda Research Institute USA, Inc., San Jose, California, USA
MARK STEYVERS, University of California, Irvine, Irvine, California, USA

As robots and autonomous agents integrate into society, understanding their influence on human social dynamics is crucial. We investigate human–robot interactions, focusing on the impact of prosocial behavior by robots on subsequent human interactions and humans' willingness to exhibit prosocial behavior toward robots. Our study involved a token-collection game in a grid-world environment. Players, human or robot, could become trapped; a prosocial action involved another player freeing the trapped individual. Findings indicate that robots demonstrating prosocial behavior toward humans can inspire prosocial behavior toward others. Humans also show a notable propensity to assist robots. Witnessing robots engage in prosocial behavior may activate social norms related to cooperation, prompting humans to emulate these behaviors. Robots' actions could improve the saliency of these acts, focusing people's attention on prosocial behaviors they might not notice otherwise. Overall, the findings suggest that robots can promote prosocial behavior among humans, contributing to a more cooperative social environment. This research has implications for design and implementation of future autonomous systems, emphasizing the importance of social considerations in human-AI interaction studies.

CCS Concepts: • **Human-centered computing** → Empirical studies in collaborative and social computing;

Additional Key Words and Phrases: Human-robot interaction, Prosocial interactions, Social robotics

ACM Reference format:
Xinyue Hu, Shashank Mehrotra, Teruhisa Misu, Kumar Akash, and Mark Steyvers. 2025. Toward Promoting Prosocial Interactions between Humans with Autonomous Agents. *ACM Trans. Hum.-Robot Interact.* 15, 1, Article 4 (August 2025), 20 pages.
https://doi.org/10.1145/3746462

1 Introduction

The proliferation of artificially intelligent agents has introduced a new dynamic into the human social environment. As autonomous agents become more common in everyday service applications, we need to consider how they can promote human well-being in the emerging hybrid society. Studies in the field of human-AI interaction often focus on trust and cooperation between humans

Authors' Contact Information: Xinyue Hu (corresponding author), University of California, Irvine, Irvine, California, USA; e-mail: xhu26@uci.edu; Shashank Mehrotra, Honda Research Institute USA, Inc., San Jose, California, USA; e-mail: shashank_mehrotra@honda-ri.com; Teruhisa Misu, Honda Research Institute USA, Inc., San Jose, California, USA; e-mail: tmisu@honda-ri.com; Kumar Akash, Honda Research Institute USA, Inc., San Jose, California, USA; e-mail: kakash@honda-ri.com; Mark Steyvers, University of California, Irvine, Irvine, California, USA; e-mail: mark.steyvers@uci.edu.

ACM 2573-9522/2025/8-ART4
https://doi.org/10.1145/3746462

and autonomous agents [45, 56]. Researchers have also focused on developing hybrid systems that surpass the respective capabilities of both humans and autonomous agents [31, 49, 53]. However, another critical aspect is the social dimension: how do social dynamics change with the inclusion of autonomous agents? [37]. People may follow social norms introduced by autonomous agents, such as conversational roles or dynamics [35, 51]. Additionally, researchers can develop policies to encourage greater acceptance of autonomous agents in the public realm [8]. This research investigates the prosocial interactions between humans and autonomous agents, and specifically, how people's prosocial behavior toward others is influenced by the actions of autonomous agents.

Research in this area has been limited in several ways. First, much of the existing research has utilized economic games to measure participants' tendency to behave prosocially toward autonomous agents such as robots [11, 20, 22, 31]. In these games, participants often interact with robots in a controlled environment, making decisions about the division of monetary rewards that reveal human's level of generosity or fairness compared to interactions with other humans [11, 20]. However, using economic games to probe reciprocity between humans and robots presents challenges because the relevance of monetary rewards to a robot is ambiguous for human participants. Second, previous research has mostly focused on direct reciprocity between one human and one autonomous agent [6, 11, 20, 22, 31]. However, consistent interactions with the same robot are unlikely in real-world scenarios. Most **Human–Robot Interactions (HRIs)** in social environments, such as roads, are ad hoc and brief. Therefore, understanding social dynamics in these contexts requires studying indirect reciprocity, which considers how people's behavior toward one agent may be influenced by their experiences with other agents.

To address these challenges, we designed a token-collecting game in a grid-world environment. Participants pursue a primary goal while having the opportunity to perform prosocial acts toward another agent. This setup allows us to explore prosocial interactions in a context more reflective of real-world scenarios, where prosocial behavior is not the primary focus but can emerge naturally from the environment. Our investigation into the prosocial dynamics between humans and robots focuses on two key research questions. First, we examine whether prosocial behavior exhibited by robots toward humans can influence subsequent social interactions among humans, potentially triggering prosocial behavior between them. The chain of interactions we investigate, also known as *upstream reciprocity*, is illustrated in Figure 1. Does a person (Person B) helped by a robot become more prosocial toward another person (Person C) in the future? The second research question explores whether humans are willing to exhibit prosocial behavior toward robots. Given that robots are not fully autonomous and may sometimes need human assistance, understanding human willingness to help robots is essential. The knowledge gained from this research will enable us to explore the potential of using autonomous agents to promote well-being among humans and develop better policies for fostering acceptance of robots in our society.

2 Related Works

2.1 Prosocial Interactions in HRI

Prosocial behavior has diverse definitions [43, 46], but at its core, it describes actions aimed at benefiting another individual [46]. Being prosocial increases the well-being of the helper and the recipient [55]. While prosocial behaviors among humans have been extensively studied, questions remain about how these interactions translate to human–robot contexts, presenting novel research opportunities [3, 25, 41, 47]. In the context of direct reciprocity, which is described as reciprocal cooperation between two individuals, the findings demonstrated what has been referred to as "AI exploitation" [11, 22, 26, 31]. In this dynamic, humans trust an autonomous agent partner to the

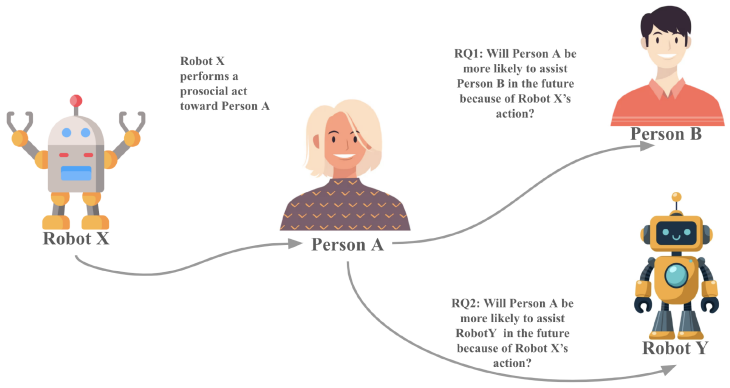


Fig. 1. An illustration of our two research questions (RQ1 and RQ2) related to different types of upstream reciprocity.

same extent that they would another human, but take more advantage of benevolent behavior when it originates from an autonomous agent than from another human.

Consider this scenario: a delivery robot is on its way to deliver an item and sees someone drop something. Picking up and returning the item is a prosocial act. If this prosocial act initiated by the delivery robot has a positive impact on the human receiving the help, that person might be more inclined to be prosocial in the future when they witness another person accidentally dropping something. The interaction discussed above is described as *upstream reciprocity*, which is the tendency to be more inclined to assist a new person if you have previously been aided by others, which is also referred to as pay-it-forward reciprocity [34, 36]. This tendency is often attributed to emotional and affective processes. Receiving help leads to feelings of gratitude, which then directly motivates the individual to give back in kind [32]. Alternative explanations for these upstream effects have centered on behavioral mimicry, wherein an individual observing assistance might mirror that behavior in interactions with others [16]. Empirical support for upstream effects comes from field studies, such as observing drivers, who are more inclined to stop for others if someone just stopped for them, in parking lots [34]. Upstream reciprocity has not been studied to the extent of direct reciprocity in HRI.

2.2 Prosocial Interactions in Economic Games

Most previous studies on the topic of prosocial interactions between human and autonomous agents have utilized economic games to examine direct reciprocity between humans and autonomous agents, treating prosocial interaction as a primary task with an explicit utility value in one-shot interactions [13, 20, 22, 26, 31, 54]. Specific examples include well-known games such as prisoner's dilemma and stag hunt [26]. When interactions extend across multiple iterations, it has been observed that autonomous agents can learn to build cooperative relationships with humans, mainly when these agents use simple non-binding signals for communication [11].

Although economic games are an effective way to study prosocial interaction between humans and autonomous agents, the setup of most economic games makes prosocial interaction the primary task of the experiment. These games are often represented in matrix form, where the rows and columns symbolize the possible strategies of the players [4]. Using these matrix-form stochastic games presents scenarios in which it is immediately obvious when another agent is helping or can benefit from prosocial actions. For instance, in one study, participants and a robot sit in front of a computer screen to play a prisoner's dilemma game, where all four options available to the

participant explicitly indicate how prosocial the participant or the robot will be by choosing each option, such as choosing to give oneself 10 coins and the robot 0 coins [20]. However, setups like these in simple economic games do not always reflect real-world social decisions, as prosocial interactions in a spatial context might not be the primary task in an environment and require awareness of the other person's actions and their need for assistance, which is not necessarily related to the primary task the agents are performing.

2.3 Prosocial Interactions in Spatio-Temporal Games

Recently, researchers have extended the scope of economic games to include stochastic games played within spatial grid environments [11, 28, 33]. Such games demand planning concerning spatial actions and thinking about the other player's intentions temporally over the course of several actions. The coordination and planning required by spatio-temporal games reflect complex social decision-making, offering valuable insights into the dynamics of human cooperation and competition. For instance, hierarchical models were developed using classic economic games in a spatio-temporal setting to establish joint intentionality [28]. A spatial variation of the Prisoner's Dilemma game was employed to demonstrate that social perceptions strongly predict preferences for artificial agents [33]. Spatial game environments bring the study of prosocial behavior closer to real-world situations in which humans and robots share the same space, such as interactions with delivery and cleaning robots. This makes the findings more relevant to understanding human interactions beyond the confines of traditional economic games. In this context, both agents—the human and the autonomous agent—might be engaged in a primary task unrelated to prosocial behavior, making any prosocial act a secondary, optional task without explicit utility. Recall the delivery robot picking up the pen in the example above: picking up the pen and returning it to the person is a prosocial act unrelated to the delivery robot's primary task. In this example and our behavioral experiments, prosocial behavior emerges from situational awareness rather than as a premeditated objective.

2.4 Indirect Reciprocity in Task-Oriented HRI

Some prior research has examined how social robots influence human behavior. One study found that when children interacted with a more prosocial robot, they were more willing to share stickers with other children [42]. Similarly, other studies on social robot behavior in children have shown that negative (antisocial) behaviors are reciprocated at an earlier age, whereas prosocial behaviors are learned later in development [7]. Additionally, when children observed prosocial or antisocial behavior from robots, they demonstrated less sharing behavior during a dictator game after witnessing antisocial behavior [40]. However, although these studies investigate the effects of robot behavior on human behavior, they do not examine adult populations, do not explore indirect reciprocity, or do not analyze prosocial behavior in a task-oriented environment where helping is not explicitly incentivized.

Additionally, some studies have examined the impact of robot behavior on adult populations. Prior research on carryover effects in HRI suggests that human behaviors toward others can be shaped by prior interactions with a robot. Even when the initial interaction was not explicitly prosocial, people adjusted their behavior toward others based on their experience with a social robot during an inclusive or exclusive ball-tossing task [14]. Similarly, people's encouragement style was influenced by their previous interaction with a robot, depending on whether the robot provided them with polite or impolite encouragement [19]. However, while these studies explore related topics, it is important to highlight the fundamental differences between our study and previous research.

While these studies provide valuable insights into how human behavior is shaped by robotic interactions, prosocial behavior in these settings was either the primary focus of the task or measured indirectly—for example, by assessing participants' physical distance from the interviewer after interacting with the robot [14, 19]. Our study extends this line of research by examining task-based, action-driven prosocial interactions in a spatial game environment, where participants must actively decide whether to engage in explicit helping behaviors. Unlike prior research that explicitly instructs or emphasizes prosocial behavior, our approach allows prosocial actions to emerge naturally as secondary tasks in an environment where helping is not the primary objective.

Moreover, as we discussed before, past research has not fully explored how indirect reciprocity functions in goal-driven environments with adult population. The economic game literature has traditionally focused on explicit, utility-based decision-making, often in one-shot or matrix-form games where prosocial actions are clearly defined [20, 22, 26, 31]. However, real-world interactions, such as those with delivery robots, occur in dynamic settings where prosocial opportunities are not explicitly framed as part of the primary task. Our study addresses this gap by investigating how participants' exposure to prosocial robot behaviors influences their willingness to assist others within a complex, task-oriented environment.

By situating our study within spatio-temporal game environments, we align with recent research that extends economic games into dynamic, spatial settings [28]. These environments better reflect real-world social interactions, where prosocial actions are often optional, effortful, and contingent on situational awareness. In sum, our research builds upon previous work by examining prosocial behavior propagation in a multi-agent environment where helping is not explicitly incentivized. By investigating upstream reciprocity toward both humans and robots, we provide new insights into how autonomous agents influence social norms in task-oriented settings and how humans generalize prosocial behaviors across different agent types.

2.5 Present Research

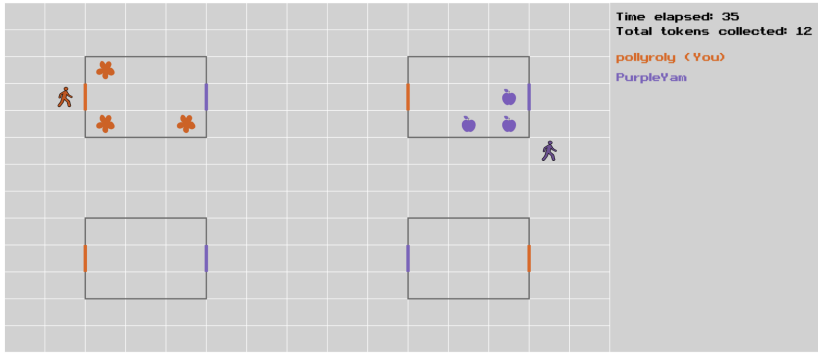
In this research, we conduct behavioral experiments to investigate upstream reciprocity between humans and autonomous agents. We are particularly interested in the tendency to be more inclined to assist a new person if one has previously been aided by others—across different agent types. Our main research question is whether people demonstrate upstream prosociality toward other humans after receiving prosocial behavior from a robot. Additionally, we investigate whether participants exhibit upstream reciprocity toward robots after being assisted by other robots.

For instance, in the scenario where an individual assists someone by picking up something they dropped on the way to a printer, the question arises whether that person will perform a similar action in the future if the assistance is initiated by a robot. Moreover, if a person is assisted by a robot, will they demonstrate a similar kind of upstream reciprocity toward another robot, considering that current applications with autonomous agents still occasionally need human assistance?

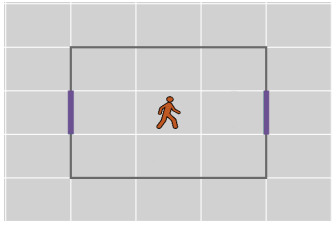
This study examines whether the type of agent—human or robot—affects the likelihood of people reciprocating or helping others in future interactions. In our experiments, these questions are addressed using both quantitative and qualitative methods. Throughout this article, we use the term "autonomous agent" to refer to entities capable of executing goal-directed behaviors independently. This includes behaviors that are pre-programmed or scripted, as long as they are perceived as intentional and self-initiated by participants. While this form of autonomy does not imply cognitive or emotional sophistication, it reflects independence commonly observed in service or delivery robots.

3 Method

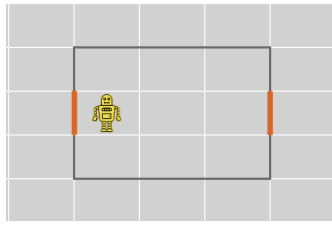
We designed a token-collecting game in a 2D grid world to simulate interactions between humans and robots in a spatial environment. In previous research using spatio-temporal variants of economic



(a) Game environment: The participant controls the orange player, who collects orange flower-shaped tokens and can only move through orange doors. The purple player, another human player, collects purple apple-shaped tokens and can only pass through purple doors. When a player enters a room to collect tokens, the door colors change, forcing them to exit the room on the other side or become trapped inside (if both doors mismatch the player color). Each round, the color and ID of the other player changes, indicating a different player.



(b) The participant's controlled player is trapped because both doors mismatch the player's color. The player can be saved by a human or robot player choosing the enter the room even though the room contains none of the other player's tokens.



(c) A trapped robot player can be saved by the participant, providing an opportunity for the participant to be prosocial.

Fig. 2. An overview of the game setup and different stages of gameplay.

games, prosocial interactions between agents were foregrounded as the primary task, and rewards for being prosocial were explicit and directly observable [20, 28, 54]. In contrast, in our experimental setup, participants' primary task of collecting tokens was not directly related to prosocial acts, and participants were neither rewarded nor punished for performing prosocial acts. The game setup is illustrated in Figure 2. Players, humans or robots, became trapped in designated areas of the game space. The prosocial action—carried out by either another human or a robot player—involved rescuing the trapped player to facilitate continued token collection.

In each round, participants were paired with a different player, either another human or a robot, to examine the effects of indirect reciprocity [2, 10]. By manipulating the sequence of agents and the opportunities for prosocial acts, we explored how these factors influenced players' prosociality across multiple rounds. If participants were paired with another human player in one of the rounds, they were playing alongside a simulated human player using replayed human movement sequences from a previous experiment rather than a real human player. We chose this replay design to maximize experimental control while ensuring that the other human player, like the robot player,

Fig. 3. Flow of the experiment: Each round lasts 90 seconds, with a total of four rounds. Participants completed an instruction phase at the beginning and an end-of-trial questionnaire at the conclusion. The entire experiment takes approximately 14 minutes to complete.

always assisted the participant when given the opportunity. As a result, the replay design ensured compatibility between the robot and other human player conditions.

The experiment was not explicitly presented as a cooperative or competitive environment. Participants received no specific instructions regarding prosocial behavior, and the game neither rewarded nor penalized such actions. Additionally, the mechanism of the prosocial act—saving a trapped player—had to be observed by participants rather than being explicitly revealed through experiment instructions or game settings. As a result, by positioning the prosocial act as an implicit secondary task within the environment, we were able to observe the reasoning behind choosing to be prosocial or not when no explicit utility was involved.

3.1 Participants

A total of 401 participants were recruited through Prolific[1] (50% male, 49% female), aged 18 to 76 (M = 38, SD = 12). All participants resided in the United States and self-reported being English speakers. Informed consent was obtained from all participants. This study received approval from the University of California, Irvine Institutional Review Board under protocol number #3624.

3.2 Procedure

3.2.1 Experimental Procedure. Participants took part in the experiment, which was hosted on a web site and consisted of an online token-collection game set in a simple grid-world environment. The experiment lasted about 14 minutes on average. Figure 3 illustrates the flow of the experimental setup. Participants were informed that the experiment's goal was to examine their strategy in a two-player interaction game. Participants first completed a tutorial that guided them through the game's setup. This tutorial included the mechanisms for collecting tokens and switching doors. The tutorial instructions made no mention of the possibility of becoming trapped in one of the rooms, nor did they specify whether the game was competitive or cooperative. After completing a game tutorial, participants were asked to assign an ID to their player, which would be used during gameplay.

Following the experiment, participants were asked to complete a post-trial questionnaire. This questionnaire sought to elicit information about the participants' interactions with the other players, as well as their motivations for helping or not helping the other player. After completing the post-trial questionnaire, participants who interacted with another human player were informed that this player was based on a replay of previously collected human movement patterns (see later section on the human replayer implementation), and they were given the option to withdraw from the study.

3.2.2 Measurements. We gathered time-series data that recorded player movement and the current game state at each of both players' moves to aid in quantitative analysis. We collected qualitative data through a post-game questionnaire. The questions focused on participants' awareness of the other player's situation (e.g., whether they noticed that the other player was trapped), their

[1] https://www.prolific.com/.

assessment of the other player's helpfulness, their perception of the game's nature (competitive or cooperative), and their reasons for helping or not helping the trapped player.

3.2.3 Gameplay Details. The experiment consisted of four rounds of gameplay, where each participant was paired with another player who could be either a robot or a human (replay player). Each round lasted 90 seconds. To simulate an online gaming experience, a 15-second countdown was initiated at the start of each round, with the message "Finding a new human player for the new round ... Trying to find another player in X seconds." It was implied that pairing with the robot player would only occur if no human players were available. At the start of each round, participants were told whether their fellow player would be a robot or a human player, as well as their ID. Participants were reminded at the start of each round that they were playing with a player who was different from the previous round.

During each round, either the participant or a fellow player would be confined within a room after a predetermined amount of gameplay, while the non-confined player had the opportunity to demonstrate prosocial behavior by entering the room to free the trapped player. Figure 2(b) and (c) illustrates this phase of the game. Players began each round diagonally across from one another, one in the bottom-right corner and the other in the top-left corner. They collected tokens of various colors and shapes, each unique to the player. Two counters in the top-right corner of the screen kept track of the time since the game's start and the total number of tokens collected by both players. Below the counters, both players' IDs were displayed, with text colored to match the colors of their icons.

Tokens were generated at random in groups of three and placed in one of the four rooms on the grid, making sure that no two players' tokens appeared in the same room. When a player finished collecting their tokens, a new set of three appeared in a different room. The participant (represented by the orange avatar) entered rooms via an orange door, whereas the other player (shown in purple for illustrative purposes) used a purple door. Each time a player entered a room, the door colors changed, teaching them how to manipulate door settings to reset room access.

Figure 2(b) and (c) depicts scenarios in which the participant or robot player is trapped in a room because both doors do not match the player's color. A trapped player can only be freed if the other (untrapped) player enters the room and resets the door colors, as they are unable to exit through either door. The intended trapped player for each round would be confined to the first room they entered after at least 20 seconds passed in the round. The timing of the trapping event was set to ensure that all participants had engaged with the game round for a sufficient duration when the trapping occurred.

3.2.4 Conditions. Table 1 list all conditions included in the experiment. We used the following notation for the conditions: the letters "H," "R," and "P" refer to the other human (replay) player, the robot player, and participant, respectively. The first four conditions (A–D) included scenarios where participants received help from another player (human or robot) during the first two rounds of the game. In the last two rounds, they were then given one opportunity in each round to help the other player (human or robot). The type of agent from whom they received help, as well as the type of agent whom they could help, remained consistent across the first two and last two rounds. For instance, in condition C, the participant is first helped by another human player in the first two rounds ($H \rightarrow P$), followed in rounds three and four by a chance to help a robot player ($P \rightarrow R$). The last two conditions (E–F) represent conditions where participants never received help from another player. In each of the four rounds of the game, they had an opportunity to help the other player (human or robot). Conditions E and F allow us to assess the baseline tendency of participants to help the other player if they were never helped themselves.

Table 1. The Six Conditions in the Experiment

Label	Condition	Participant Count
A	$R \rightarrow P, P \rightarrow H$	69
B	$H \rightarrow P, P \rightarrow H$	61
C	$H \rightarrow P, P \rightarrow R$	68
D	$R \rightarrow P, P \rightarrow R$	66
E	$P \rightarrow H$	67
F	$P \rightarrow R$	70

P = participant; R = AI robot player; H = another human player. The arrow represents the direction of help that can be offered. For example, $P \rightarrow R$ represents a scenario where the participant has an opportunity to help the robot player.

A between-subject design was employed in the experiment. Participants were randomly assigned to one of the six experiment conditions. Table 1 denotes the count of participants in each condition. To reinforce the perception of interacting with different players in each round, we varied the color of the other player's icon, player ID, and tokens. The participants were represented by an orange avatar and collected orange flowers throughout the game, while the colors for the other human player and robot player changed after each round. Colors were assigned so that in each of the four rounds, a unique color was used for the other player in every round. Participants were represented by an orange avatar, which remained consistent throughout the experiment. Consequently, in each round, the other player had a different color, ensuring that there were no repeated colors within the same round, and no player had the same color as the participant. The other human player collected apple-shaped tokens, and the robot player collected butterfly-shaped tokens. In total, four colors were used in the experiment (purple, green, blue, and yellow). The player IDs for the replay players were randomly generated, and the player IDs for the robot players were given generic names such as "YellowRobot2."

3.2.5 Robot Player Implementation. The robot player was implemented using an A* path-finding algorithm and moved at a constant speed of three grid positions per second. Based on a pilot experiment, we determined that this speed is comparable to the average speed of human players. The robot player's movement was divided into four stages: navigating to a room, collecting tokens, rescuing the trapped player, and becoming trapped. The robot player alternated between these stages depending on the game state as shown in Figure 4. At each stage, the A* path-finding algorithm created a path from one door to another. The robot player was designed to exhibit prosocial behavior and always free the participant consistently. It began navigating toward the participant 5 seconds after they were trapped. This design feature was intended to maximize human reciprocity toward the robot player. The robot's actions were generated via scripted path-planning algorithm. While these behaviors were not responsive in real time, they were executed autonomously during gameplay, providing participants with a consistent and independently acting agent.

3.2.6 Human Replay Implementation. Some participants were paired with another human player in one of the rounds. However, they were actually playing alongside a simulated human player that used replayed human movement sequences from a previous experiment instead of a real human player. The movement data to construct the other human player were based on data from 343 participants collected in a previous pilot experiment [21]. To create a representative sample of human movements, we selected only those participants who had an average speed of 2.5–3.5 grid

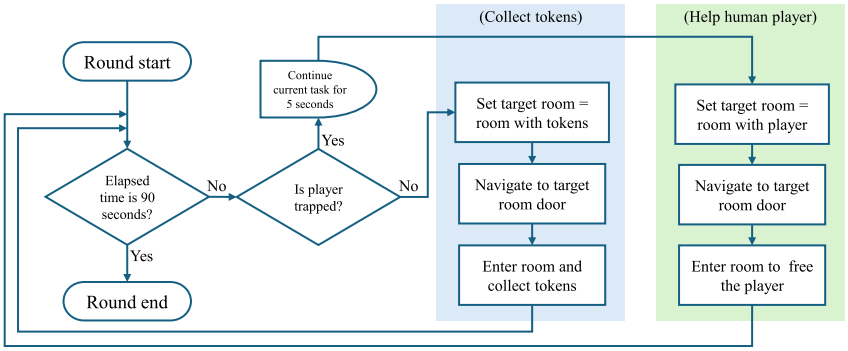


Fig. 4. An illustration of the different states the autonomous agents switch between during the experiment.

positions per second, around the mode of participants' speeds. The replay player was implemented using the same logic as the robot player. However, instead of using a path-finding algorithm to generate movement sequences, we utilized the movement and timing sequences from these subset of previous participants. For each door-to-door combination on the grid, we randomly sampled 20 paths from these participants. During gameplay, when the replay player traveled from one door to another, one of the 20 sampled paths was randomly selected, and the algorithm replayed the movement and timing sequence of that chosen path. Like the robot player, the replay player began navigating toward the participant 5 seconds after they became trapped to ensure that the perceived prosociality was consistent between different agent types.

3.2.7 Differences between Robot and Human Player. The other human and robot players were distinguished by not only their avatars but also their movement patterns. While the robot player's average speed was set to match that of the human players, the robot moved at a constant velocity with each step, whereas replayed human players had the tendency to accelerate on straight paths and slowed down at turns. In addition, the robot player tended to take more turns overall (the A* pathfinding algorithm used to plan the robot paths did not induce penalties for turns), whereas human movement was characterized by a minimal number of turns. Additionally, the robot player was optimal as it followed the calculation of A* algorithms, whereas the human (replayed) player repeated behavior from previous participants, and thus, exhibited imperfect behavior as people sometimes overshoot or miscalculate the path to take. Overall, this resulted in a noticeably different movement pattern for the robot and other human player. When the robot player was trapped, it repeated a simple movement sequence within the room, mimicking the token-collection behavior typical for that space. Similarly, the human (replayed) player exhibited the same behavior, following recorded movement patterns of token collection in the same room. This design ensured that the behavior of both agent types (robot and human) was highly similar while trapped, and neither agent remained stationary.

3.3 Statistical Analysis

We employed the Bayesian logistic regression models and Bayesian A/B tests using the JASP software packages [24] to determine the factors influencing participants' decisions to engage in prosocial behavior. In the Bayesian analyses, a Bayes factor, BF_{10}, determines the extent to which the observed data adjust our belief in the alternative hypothesis over the null hypothesis. Values of $3 < BF_{10} < 10$ and $BF_{10} > 10$ indicate moderate and strong evidence against the null hypothesis, respectively. Similarly, values of $BF_{10} < 1$ indicate support in favor of the null hypothesis.

Table 2. Percentage of Prosocial Behavior by the Participant across Conditions and Rounds

		Round number			
Label	Condition	1	2	3	4
A	$R \rightarrow P, P \rightarrow H$	-	-	66.7%	73.9%
B	$H \rightarrow P, P \rightarrow H$	-	-	67.2%	68.9%
C	$H \rightarrow P, P \rightarrow R$	-	-	51.5%	64.7%
D	$R \rightarrow P, P \rightarrow R$	-	-	48.5%	54.5%
E	$P \rightarrow H$	16.4%	26.9%	34.3%	37.3%
F	$P \rightarrow R$	25.7%	28.6%	32.9%	35.7%

Empty cells indicate rounds where participant did not have an opportunity to exhibit prosocial behavior.

4 Results

In this section, we present the findings from our analyses on the propagation of prosocial behavior among participants following interactions with different agent types (human or robot). Two key research questions were explored: Do participants pass on prosocial behavior to another human when assisted by a robot, and do they exhibit similar behavior toward robots after receiving help from one? We begin by reporting the extent to which receiving assistance from an agent (human or robot) influences participants' subsequent likelihood of assisting another agent. We then describe participants' stated motivations for engaging or not engaging in prosocial behavior. Finally, we examine how the similarity or difference in agent type (human vs. robot) affects these motivations. All data collected from this study can be accessed here.

4.1 Participants Equally Pass on Prosocial Behavior to Humans When Helped by Humans or Robots

Table 2 shows the proportion of prosocial acts by the participant by condition and round. The Bayesian A/B test comparing conditions A and E provides strong evidence that participants are significantly more inclined to assist another human after being aided by a robot, compared to having received no prior help ($BF_{10} > 100$, CI = [1.20, 2.04]). Additionally, there is no evidence to suggest that participants' tendency to pass forward prosocial behavior to another human differs based on whether they were helped by another human or a robot ($BF_{10} < 1$) when comparing conditions A and B. Overall, participants demonstrate upstream reciprocity toward another human equally, regardless of the type of agent that previously assisted them.

Furthermore, when examining the first four conditions (A–D) in Table 2, there is no evidence to suggest that the type of agent showing prosocial behavior toward participants influenced their decision to pass on the prosocial act, regardless of the type of agent receiving the help ($BF_{10} < 1$). This result indicates that in this context robots have a similar level of social impact on human behavior, as participants are equally likely to be prosocial toward another agent in the future after being helped by a robot compared to being helped by a human. Additionally, when examining whether there is an impact on participants' tendency to be prosocial based on whether the type of agent who previously helped them and the type of agent they have the opportunity to help are the same, the results indicate that there is no effect of intergroup versus intragroup interactions. This suggests that prosocial behavior in the context of this study is influenced not only by the identity of the helper or the recipient but also by the act of receiving the prosocial behavior itself.

4.2 Participants Pass on Prosocial Behavior to Robots, but Less Likely than to Humans

When comparing conditions D and F as shown in Table 2, there is strong evidence that participants are more likely to help another robot after being helped by a robot, compared to when they have never received help ($BF_{10} > 100$, CI = [0.42, 1.25]). Additionally, the Bayesian A/B test comparing conditions C and D shows no evidence that participants' tendency to pass forward prosocial behavior to a robot differs based on the type of agent that helped them ($BF_{10} < 1$). Moreover, there is strong evidence that participants are less likely to perform a prosocial act toward a robot player compared to a human player ($BF_{10} > 10$, CI = [−0.762, 0]), which aligns with previous research indicating that humans do not exhibit the same extent of prosociality toward robots as they do toward humans [22]. Overall, participants exhibit upstream prosociality toward other robots after being helped by robots.

4.3 Increased Prosociality over Time

The Bayesian logistic regression model with all six conditions shown in Table 2 revealed that the more rounds participants interacted with other players, the more likely they were to save the trapped player ($BF_{10} = 82.8$, CI = [−0.04, 0.45]). The effects of rounds suggest that the more time participants spent interacting with another player, the more salient the other player's actions became. Consequently, participants were more likely to be prosocial toward the other player as they became more aware of the other player's situation.

4.4 Exploratory Qualitative Analysis

For the qualitative analyses, three researchers independently coded the responses for the reasons why participants chose to help or not help. The coding scheme was created by iterating through the emergent themes based on the reasons given for helping or not helping the robots. The approach for developing the qualitative codes followed principles of Grounded theory [12]. The emergent qualitative themes were representative of the reasoning for why participants helped or did not help. Table 3 displays the top five motivations for both helping and not helping. We conducted a logistic regression model on each motivation to examine the impact of different factors on participants' motivations to help or not help another agent. These factors include the type of agent receiving help, the type of agent previously assisting the participant, whether or not the participants have previously received help, and whether the types of agents receiving help and assisting are the same.

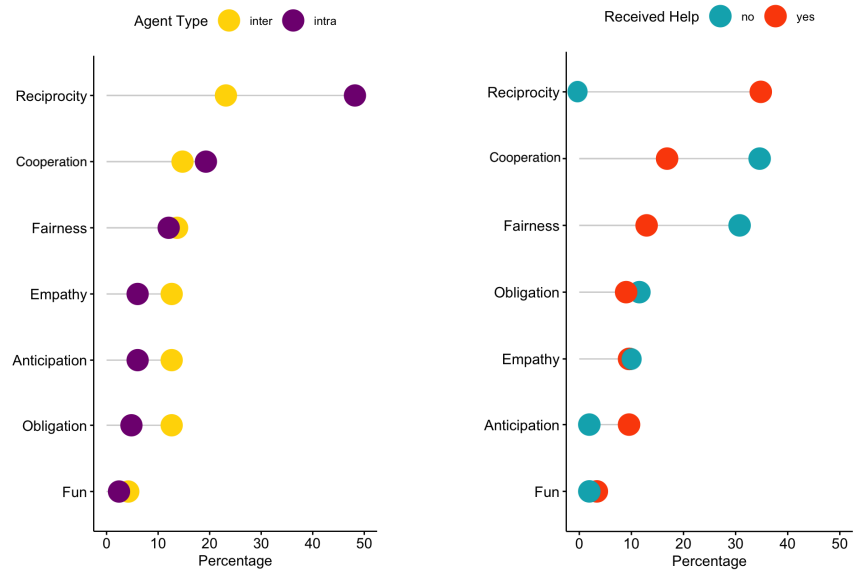
4.4.1 Motivation to Help. When comparing the impact of each predictor on each motivation, we found no evidence suggesting that the type of agents assisting the participants before, or the type of agents they assist, leads to a difference in the likelihood of any reported motivations ($BF_{10} < 1$). This indicates that, in general, participants' motivations to help are not significantly influenced by whether they previously received assistance from a human or a robot, nor by the type of agent they are assisting.

However, we found strong evidence that people's motivation is more likely to be reciprocity when the types of agents they receive help from and help are the same, compared to when they are different types of agents ($BF_{10} = 39.9$, CI = [0.35, 1.77]), as shown in Figure 5(a). This suggests a specific and notable pattern: when participants both receive and provide help to the same type of agent, they are significantly more likely to cite reciprocity as their motivation. This finding highlights the role of perceived continuity in fostering reciprocal behaviors. When the agent types differ, this sense of reciprocity decreases, possibly due to perceived differences between agent identity. No other motivations are influenced by this factor ($BF_{10} < 1$).

Table 3. Frequency of Top Five Themes for Saving the Other Player and Top Five Themes for Not Saving the Other Player

Themes	Sample Statement	Percentage
Motivation to help:		
Reciprocity	"I was helping because I felt super bad, after the others helped me"	27% (62)
Cooperation	"I wanted my partner to be able to achieve as many points as possible"	21% (48)
Sense of fairness	"It felt bad that they were not getting a fair chance to beat me"	17% (39)
Empathy	"The first time I didn't but the second time I felt bad and helped"	10% (22)
Obligation	"Because it was the right thing to do"	10% (22)
Motivation to not help:		
Unawareness	"Because I didn't notice it"	32% (55)
Competition	"I thought it was a competition"	26% (45)
Prioritization	"I wanted to get more tokens"	20% (35)
Inability	"I didn't know I could help"	13% (22)
Indifference	"I didn't care that he was stuck"	8% (13)

The number of participants is shown in parentheses. This table includes all six conditions.



(a) Motivations to help another player by agent type.

(b) Motivations to help another player based on prior receipt of help.

Fig. 5. Motivations to help another player across (a) agent type and (b) prior receipt of help. The x-axis represents the percentage of each motivation among all participants who helped either a human player or a robot player. The y-axis represents the specific motivations for helping. For (a), inter-agent refers to helping a different type of agent, while intra-agent refers to helping the same type of agent, based on the type of agent that helped the participants. For (b), "received help: yes" means that the participants were helped by another agent previously, and "received help: no" means that the participants were never helped by another agent.

Additionally, reciprocity, cooperation, and fairness are more likely to be participants' motivations to help when they have received help from another agent, compared to when they have never received help from another player ($BF_{10} > 100$, $BF_{10} = 3.9$, and $BF_{10} = 6.2$), as shown in Figure 5(b). It is worth noting that reciprocity shows a significant difference depending on whether participants have received help or not, because none of the participants would explain their choice to help as a form of reciprocity if they have never been helped before.

4.4.2 Motivation to Not Help. There is no evidence indicating that any of the participants' motivations for not helping differ based on any of the predictors described above ($BF_{10} < 1$). Participants demonstrate similar patterns across different agent types and previous experiences. As shown in Table 3, unawareness and competition are the two main motivations behind participants' decisions to not help another agent. Here, unawareness refers to participants who responded with comments such as "I may have just missed it by being hyper-focused on the flowers," indicating that they either did not notice or were unaware of the robot's situation due to being preoccupied, and competition refers to the motivation to outperform the other agent.

4.5 Results after Excluding Participants Reporting Unawareness or Inability to Help

A portion of participants reported that their motivation for not helping was due to being unaware that another agent was trapped and needed assistance (55 participants). Additionally, some participants (22) indicated that they were unsure of the game mechanics required to free a trapped player from a locked room. This may not accurately reflect group-level behavior, as participants' failure to help was not necessarily due to unwillingness to engage in prosocial behavior but rather their inability to understand the game mechanics or recognize the opportunity to assist another agent. The detailed summary of prosocial rate among participants is provided in Appendix A.

To address this, we created a subset of the data by excluding participants who reported unawareness or inability as their motivations for not helping and who never helped other agents during the experiment (45 participants). This adjustment resulted in a final subset of 356 participants. The same data analysis as in the previous section was conducted, and the results are reported below. Table A1 in Appendix A shows the frequency of prosocial behavior in each opportunity for each condition.

Here, we observe a similar pattern in participants' behavior in the subset of data compared to the full dataset, as shown in Sections 4.1–4.3. When examining the first four conditions (A–D), there is no evidence to suggest that the type of agent demonstrating prosocial behavior toward participants influenced their decision to pass on the prosocial act, regardless of the type of agent receiving the help ($BF_{10} < 1$). Additionally, there is no effect of time or intra-agent vs. inter-agent type.

When comparing conditions A and E using the Bayesian A/B test, we observe the same pattern: participants are significantly more inclined to assist another human after being aided by a robot compared to having received no prior help ($BF_{10} > 100$, CI = [0.98, 1.90]). Participants' tendency to pass forward prosocial behavior to another human does not differ based on whether they were helped by another human or a robot ($BF_{10} < 1$) when comparing conditions A and B. As before, participants demonstrate upstream reciprocity toward another human equally, regardless of the type of agent that previously assisted them.

As before, the Bayesian logistic regression model including all six conditions shown in Table A1 revealed that the more rounds participants interacted with other players, the more likely they were to save the trapped player ($BF_{10} = 14.7$, CI = [−0.01, 0.56]). These findings suggest that prolonged interaction with other players made their actions more salient to the participants.

When examining participants' prosocial behavior toward robots by comparing conditions D and F, we again find that participants are more likely to help another robot after being helped by a robot,

compared to when they have never received help ($BF_{10} > 100$, CI = [0.57, 1.48]). The Bayesian A/B test comparing conditions C and D indicates that participants' tendency to pass forward prosocial behavior to a robot does not differ based on the type of agent that helped them ($BF_{10} < 1$). However, participants are less likely to behave prosocially toward a robot player compared to a human player when conditions A and D are compared ($BF_{10} = 7.50$, CI = [−0.71, 0]).

5 Discussion

As robots and other autonomous agents become more integrated into society, it is important to understand their impact on human prosocial behavior. Most previous research on prosocial interaction between humans and robots has focused on direct reciprocity between one participant and one robot [26, 27, 39, 42, 58]. In addition, some previous studies in HRI have indicated that emotion expression or actions initiated by robots can influence participants to be more prosocial toward another robot [9, 15, 57]. In contrast, our research focuses on the broader question of how robot prosociality toward humans encourages people to be prosocial toward other humans who are not limited to cooperation or sharing behavior [42]. The findings provide strong empirical evidence of robots' social influence on human behavior toward one another. Our findings show that participants exhibit strong upstream reciprocity toward other humans regardless of whether they received prosocial acts from a robot or another human. Taken together, our findings show that robots can be used to promote human well-being.

One explanation for the increase in prosocial behavior, when participants received assistance from robots as opposed to when they did not receive assistance, is that the robots' behavior served as a model for participants on how agents should behave. The robots' prosocial actions could be interpreted as establishing a norm in the environment that participants believed that they were expected to follow. For future projects investigating the possibility of using autonomous agents to establish societal norms, researchers could investigate whether people imitate robot actions in environments with various types of actions, such as negative actions, neutral actions, and other types of prosocial actions.

Several studies have been conducted on social norms and dynamics among human agents [5, 18, 50]. Nonetheless, there remains a gap in the literature regarding the role of autonomous agents in establishing positive social norms in society. To understand how we can use autonomous agents to promote positive social norms, we must first understand the nature of prosocial interactions involving robots. For example, if assisting a human in a given situation is considered a moral norm, where you feel obligated to help another person, would introduce a robot into the social dynamic change it to a prudential norm, where being prosocial is not required but is seen as an honorable thing to do [23]? Future research could address these types of questions by conducting a more in-depth qualitative analysis of the participants' reports from these interactions.

Another explanation for the increase in human prosocial behavior following exposure to robot prosocial behavior is that some participants may have been unaware of the possibility of prosocial behavior. In the context of our experiment, helpful behavior consisted of freeing a trapped agent from a room. Some participants may not have realized that entering the room could free another agent. Consistent with this explanation, our findings indicate that unawareness is the most important reason for not assisting the other agent. In addition, participants became more prosocial in the later rounds compared to the early rounds in conditions E and F. Participants' remarks such as "In the first round, I did not notice they were stuck. Then, after I noticed, I helped after some time passed," highlighting that participants were unaware of the possible ways to help other agents or unaware of the other player's situation. As time passed, the other player's actions and situations became more salient, thus increasing the likelihood of participants' prosocial acts. This finding emphasizes the difference between spatial-temporal games, such as the one we designed, and economic games.

In economic games, the social dynamics between the agents are the primary task in the experiment and the focus of participants' attention [11, 20, 22, 26, 31]. In our game, the participants' main task is to collect tokens, and they do not necessarily need to pay attention to the other players. As a result, participants must actively consider the other player's actions and situations to decide to save a trapped player from a closed room. Our results highlight an important aspect of prosocial behavior that is typically not discussed in the relevant literature: the role of attention. To engage in prosocial interactions, participants must be aware of the other agent's situation. This results suggest that robots' actions could act as reminders in the environment, highlighting possible prosocial acts and improving the saliency of these actions.

To ensure that our findings were not solely driven by participants who were unaware of the possibility of prosocial behavior or unable to perform prosocial actions due to difficulty understanding the saving mechanism, we conducted an additional analysis excluding participants who explicitly reported "unawareness" or "inability" as their reason for not helping. Importantly, even after removing these participants, the core pattern of prosocial behavior remained consistent, indicating that the observed upstream reciprocity effect was robust. Additionally, our prior research has demonstrated that awareness plays a critical role in prosocial decision-making, particularly in spatial environments where prosocial opportunities are not the primary task, reflecting a more ecologically valid representation of everyday social interactions [21].

While our findings suggest that robot behavior has an influence on human prosociality, it is important to recognize the simplicity of the autonomous agents in our study. There is no dynamic social responsiveness or emotional expression during gameplay that was expressed by the robots. Rather than engaging in real-time interactions, participants encountered consistent, scripted behaviors that may not have been perceived as intentional social acts. However, despite this simplicity, some participants still described the robot's actions in explicitly social and moral terms, as they expressed feelings of fairness, empathy, or guilt, as shown in Section 4.4. These reports suggest that even minimal behavioral cues from autonomous agents can elicit social interpretations and motivate prosocial action. Future studies could explore how emotionally expressive and responsive agents shape perceptions of social intentionality in HRIs.

6 Conclusion

The findings presented above highlight the dual role of autonomous agents—not only as facilitators of prosocial behavior but also as catalysts for increasing the salience of prosocial opportunities in dynamic, task-oriented environments. This insight has important implications for the design and policy considerations of autonomous agents, suggesting that their presence can be leveraged to promote social cooperation in real-world settings.

It is worth noting that in this project, we employed a virtual representation of the robot instead of a physically embodied robot, as is commonly used in most HRI studies. The differences between these two types of representations have been widely discussed in previous research [17, 29, 30, 38, 52]. The physical presence of a robot, compared to one displayed solely on a screen, provides people with a greater sense of social presence, with physically embodied robots generally eliciting higher levels of social presence [29]. A similar effect has been observed in interactions with children during motor tasks—both virtual and embodied robots were effective in introducing the task, but children engaged less with the virtual agent [17]. Other studies have also highlighted that physically present robots are generally preferred over virtual agents [30, 52]. However, while there are differences in the degree of social presence between virtual and physically embodied robots, both representations have been shown to effectively achieve their experimental objectives—whether it be introducing children to motor tasks or acting as proctors in exams [1, 17].

Similarly, interactions with virtual agents in video game environments have been found to influence players' altruistic tendencies [38]. Other studies have shown that virtual agents can elicit comparable social and behavioral effects as physically embodied robots in both educational and motor learning contexts [44, 48]. Of course, there are limitations when generalizing experimental results from virtual agents to physically embodied robots. However, the statistical power provided by web-based experiments, combined with the comparable impact observed between virtual and physical agents, makes this study a valuable starting point for developing empathetic agents capable of influencing prosocial behavior.

Overall, our findings reveal that participants exhibited upstream reciprocity toward both other humans and robots after being previously assisted by robots. This study contributes to the expanding field of HRI, offering a glimpse of how technology could be integrated into our established social dynamics. It underscores the potential for autonomous agents to actively contribute to our collective well-being by demonstrating how robots can influence human prosocial behavior. This research opens new avenues for exploring how technology can foster positive social interactions. As we move toward a world increasingly populated by intelligent machines, understanding and leveraging these dynamics will be crucial for building harmonious and mutually beneficial human–robot relationships.

Data Availability

All behavioral data collected for this study are publicly available from the following OSF repository: https://osf.io/5ptu8/?view_only=0748e136f5ad42b1ba96f1dff161a0d5.

References

[1] Muneeb I. Ahmad and Reem Refik. 2022. No chit chat! A warning from a physical versus virtual robot invigilator: Which matters most? *Frontiers in Robotics and AI* 9 (2022), 908013. DOI: https://doi.org/10.3389/frobt.2022.908013

[2] Hannelore Brandt and Karl Sigmund. 2004. The logic of reprobation: Assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology* 231, 4 (2004), 475–486. DOI: https://doi.org/10.1016/j.jtbi.2004.06.032

[3] Hilmar Brohmer, Andreas Fauler, Caroline Floto, Ursula Athenstaedt, Gayannée Kedia, Lisa V. Eckerstorfer, and Katja Corcoran. 2019. Inspired to lend a hand? Attempts to elicit prosocial behavior through goal contagion. *Frontiers in Psychology* 10 (2019), 545. DOI: https://doi.org/10.3389/fpsyg.2019.00545

[4] Colin F. Camerer. 2003. Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences* 7, 5 (2003), 225–231. DOI: https://doi.org/10.1016/S1364-6613(03)00094-9

[5] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. 2018. Experimental evidence for tipping points in social convention. *Science* 360, 6393 (2018), 1116–1119. DOI: https://doi.org/10.1126/science.aas8827

[6] Nadia Chernyak and Heather E. Gary. 2016. Children's cognitive and behavioral reactions to an autonomous versus controlled social robot dog. *Early Education and Development* 27, 8 (2016), 1175–1189. DOI: https://doi.org/10.1080/10409289.2016.1158611

[7] Nadia Chernyak, Kristin L. Leimgruber, Yarrow C. Dunham, Jingshi Hu, and Peter R. Blake. 2019. Paying back people who harmed Us but not people who helped us: Direct negative reciprocity precedes direct positive reciprocity in early development. *Psychological Science* 30, 9 (2019), 1273–1286. DOI: https://doi.org/10.1177/0956797619854975

[8] Samuel Chng and Lynette Cheah. 2020. Understanding autonomous road public transport acceptance: A study of Singapore. *Sustainability* 12, 12 (2020), 4974. DOI: https://doi.org/10.3390/su12124974

[9] Joe Connolly, Viola Mocz, Nicole Salomons, Joseph Valdez, Nathan Tsoi, Brian Scassellati, and Marynel Vázquez. 2020. Prompting prosocial human interventions in response to robot mistreatment. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. ACM, New York, NY, 211–220. DOI: https://doi.org/10.1145/3319502.3374781

[10] Arianna Costantini, Andrea Scalco, Riccardo Sartori, Elena Tur, and Andrea Ceschi. 2019. Theories for computing prosocial behavior. *Nonlinear Dynamics, Psychology, and Life Sciences* 23, 2 (Apr. 2019), 297–313.

[11] Jacob W. Crandall, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A. Goodrich, and Iyad Rahwan. 2018. Cooperating with machines. *Nature Communications* 9, 1 (Jan. 2018), 233. DOI: https://doi.org/10.1038/s41467-017-02597-8

[12] John W. Creswell and Cheryl N. Poth. 2016. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Sage Publications.

[13] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, K. Larson, and Thore Graepel. 2020. Open problems in cooperative AI. arXiv:2012.08630. Retrieved from https://arxiv.org/abs/2012.08630

[14] Hadas Erel, Elior Carsenti, and Oren Zuckerman. 2022. A carryover effect in HRI: Beyond direct social effects in human-robot interaction. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 342–352. DOI: https://doi.org/10.1109/HRI53351.2022.9889554

[15] Hadas Erel, Marynel Vázquez, Sarah Sebo, Nicole Salomons, Sarah Gillet, and Brian Scassellati. 2024. RoSI: A model for predicting robot social influence. *ACM Transactions on Human-Robot Interaction* 13, 2 (Jun. 2024), Article 18, 1–22. DOI: https://doi.org/10.1145/3641515

[16] James H. Fowler and Nicholas A. Christakis. 2010. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences* 107, 12 (2010), 5334–5338.

[17] Marina Fridin and Mark Belokopytov. 2014. Embodied robot versus virtual agent: Involvement of preschool children in motor task performance. *International Journal of Human-Computer Interaction* 30, 6 (2014), 459–469. DOI: https://doi.org/10.1080/10447318.2014.888500

[18] Robert Hawkins, Noah Goodman, and Robert Goldstone. 2018. The emergence of social norms and conventions. *Trends in Cognitive Sciences* 23, 2 (Dec. 2018), 158–169. DOI: https://doi.org/10.1016/j.tics.2018.11.003

[19] Kana Higashino, Mitsuhiko Kimoto, Takamasa Iio, Katsunori Shimohara, and Masahiro Shiomi. 2023. Is politeness better than impoliteness? Comparisons of robot's encouragement effects toward performance, moods, and propagation. *International Journal of Social Robotics* 15, 5 (2023), 717–729.

[20] Te-Yi Hsieh, Bishakha Chaudhury, and Emily S. Cross. 2023. Human–robot cooperation in economic games: People show strong reciprocity but conditional prosociality toward robots. *International Journal of Social Robotics* 15, 5 (2023), 791–805. DOI: https://doi.org/10.1007/s12369-023-00981-7

[21] Xinyue Hu, Kumar Akash, Shashank Mehrotra, Teruhisa Misu, and Mark Steyvers. 2024. Prosocial acts towards AI shaped by reciprocation and awareness. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*. Retrieved from https://escholarship.org/uc/item/14q5z5bv

[22] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521. DOI: https://doi.org/10.1038/s42256-019-0113-5

[23] William James. 1956. *The Will to Believe and Other Essays in Popular Philosophy*. Dover Publications, New York, 1–31. Original work published 1896.

[24] JASP Team. 2024. JASP (Version 0.18.3) [Computer software]. Retrieved from https://jasp-stats.org/

[25] Haesung Jung, Eunjin Seo, Eunjoo Han, Marlone Henderson, and Erika Patall. 2020. Prosocial modeling: A meta-analytic review and synthesis. *Psychological Bulletin* 146, 8 (May 2020), 635–663. DOI: https://doi.org/10.1037/bul0000235

[26] Jurgis Karpus, Adrian Krüger, Julia Tovar Verba, Bahador Bahrami, and Ophelia Deroy. 2021. Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience* 24, 6 (2021), 102679. DOI: https://doi.org/10.1016/j.isci.2021.102679

[27] Ran Hee Kim, Yeop Moon, Jung Ju Choi, and Sonya S. Kwak. 2014. The effect of robot appearance types on motivating donation. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI '14)*. ACM, New York, NY, 210–211. DOI: https://doi.org/10.1145/2559636.2563685

[28] Max Kleiman-Weiner, Mark K. Ho, Joseph L. Austerweil, Michael L. Littman, and Joshua B. Tenenbaum. 2016. Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society,* 38.

[29] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies* 64, 10 (2006), 962–973. DOI: https://doi.org/10.1016/j.ijhcs.2006.05.002

[30] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37. DOI: https://doi.org/10.1016/j.ijhcs.2015.01.001

[31] Kinga Makovi, Anahit Sargsyan, Wendi Li, Jean-François Bonnefon, and Talal Rahwan. 2023. Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications* 14, 1 (2023), 3108. DOI: https://doi.org/10.1038/s41467-023-38592-5

[32] Michael E. McCullough, Shelley D. Kilpatrick, Robert A. Emmons, and David B. Larson. 2001. Is gratitude a moral affect? *Psychological Bulletin* 127, 2 (2001), 249–266. DOI: https://doi.org/10.1037/0033-2909.127.2.249

[33] Kevin R. McKee, Xuechunzi Bai, and Susan T. Fiske. 2024. Warmth and competence in human-agent cooperation. *Autonomous Agents and Multi-Agent Systems* 38, 1 (2024), 23. DOI: https://doi.org/10.1007/s10458-024-09649-6

[34] Redzo Mujcic and Andreas Leibbrandt. 2018. Indirect reciprocity and prosocial behaviour: Evidence from a natural field experiment. *The Economic Journal* 128, 611 (2018), 1683–1699. DOI: https://doi.org/10.1111/ecoj.12474

[35] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction (HRI '09)*. ACM, New York, NY, 61–68. DOI: https://doi.org/10.1145/1514095.1514109

[36] Martin A. Nowak and Sébastien Roch. 2007. Upstream reciprocity and the evolution of gratitude. *Proceedings: Biological Sciences* 274, 1610 (2007), 605–609. Retrieved from http://www.jstor.org/stable/25223822

[37] Raquel Oliveira, Patricia Arriaga, Fernando Santos, Samuel Mascarenhas, and Ana Paiva. 2020. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior* 114 (Sep. 2020), 106547. DOI: https://doi.org/10.1016/j.chb.2020.106547

[38] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (Sep. 2017), Article 11, 1–40. DOI: https://doi.org/10.1145/2912150

[39] Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering pro-sociality with autonomous agents. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 1. DOI: https://doi.org/10.1609/aaai.v32i1.12215

[40] Yi Pang and Hui Li. 2024. When the recipient is a social robot: The impact of negative behavioral valence on 5-year-old children's sharing. *International Journal of Human–Computer Interaction* 41, 10 (2024), 6403–6412. DOI: https://doi.org/10.1080/10447318.2024.2378247

[41] Louis A. Penner, John F. Dovidio, Jane A. Piliavin, and David A. Schroeder. 2005. Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology* 56 (2005), 365–392. DOI: https://doi.org/10.1146/annurev.psych.56.091103.070141

[42] Jochen Peter, Rinaldo Kühne, and Àlex Barco. 2021. Can social robots affect children's prosocial behavior? An experimental study on prosocial robot models. *Computers in Human Behavior* 120 (2021), 106712. DOI: https://doi.org/10.1016/j.chb.2021.106712

[43] Stefan Pfattheicher, Yngwie Nielsen, and Isabel Thielmann. 2021. Prosocial behavior and altruism: A review of concepts and definitions. *Current Opinion in Psychology* 44 (Aug. 2021), 124–129. DOI: https://doi.org/10.1016/j.copsyc.2021.08.021

[44] Astrid M. Rosenthal-von der Pütten, Carolin Straßmann, and Nicole C. Krämer. 2016. Robots or agents–neither helps you more or less during second language acquisition: Experimental study on the effects of embodiment and type of speech output on evaluation and alignment. In *Proceedings of the 16th International Conference on Intelligent Virtual Agents (IVA '16)*. Springer International Publishing, 85–95. DOI: https://doi.org/10.1007/978-3-319-47665-0_23

[45] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2020. How social robots influence people's trust in critical situations. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1020–1025. DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223471

[46] David A. Schroeder and William G. Graziano. 2015. *The Field of Prosocial Behavior: An Introduction and Overview*. Oxford Academic. DOI: https://doi.org/10.1093/oxfordhb/9780195399813.013.32

[47] Youngsoo Shin and Jinwoo Kim. 2018. Data-centered persuasion: Nudging user's prosocial behavior and designing social innovation. *Computers in Human Behavior* 80 (2018), 168–178. DOI: https://doi.org/10.1016/j.chb.2017.11.009

[48] Masahiro Shiomi, Soto Okumura, Mitsuhiko Kimoto, Takamasa Iio, and Katsunori Shimohara. 2020. Two is better than one: Social rewards from two agents enhance offline improvements in motor skills more than single agent. *PLoS One* 15, 11 (2020), e0240622. DOI: https://doi.org/10.1371/journal.pone.0240622

[49] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119. DOI: https://doi.org/10.1073/pnas.2111547119

[50] Margaret E. Tankard and Elizabeth Levy Paluck. 2016. Norm perception as a vehicle for social change. *Social Issues and Policy Review* 10, 1 (2016), 181–211. DOI: https://doi.org/10.1111/sipr.12022

[51] Hamish Tennent, Solace Shen, and Malte Jung. 2019. Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 133–142. DOI: https://doi.org/10.1109/HRI.2019.8673013

[52] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. 2016. Physical vs. virtual agent embodiment and effects on social interaction. In *Intelligent Virtual Agents*. David Traum, William Swartout, Peter Khooshabeh, Stefan Kopp, Stefan Scherer, and Anton Leuski (Eds.). Springer International Publishing, Cham, 412–415. DOI: https://doi.org/10.1007/978-3-319-47665-0_44

[53] Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, and Yaser P. Fallah. 2021. Cooperative autonomous vehicles that sympathize with human drivers. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, 4517–4524. DOI: https://doi.org/10.1109/IROS51168.2021.9636151

[54] Wenshuo Wang, Letian Wang, Chengyuan Zhang, Changliu Liu, and Lijun Sun. 2022. Social interactions for autonomous driving: A review and perspectives. *Foundations and Trends® in Robotics* 10, 3–4 (2022), 198–376. DOI: https://doi.org/10.1561/2300000078

[55] Netta Weinstein and Richard M. Ryan. 2010. When helping helps: Autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of Personality and Social Psychology* 98, 2 (2010), 222–244. Retrieved from https://api.semanticscholar.org/CorpusID:33123603

[56] Jane Wu, Erin Paeng, Kari Linder, Piercarlo Valdesolo, and James C. Boerkoel. 2016. Trust and cooperation in human-robot decision making. In *AAAI Fall Symposia*. Retrieved from https://api.semanticscholar.org/CorpusID:37068728

[57] Shujie Zhou and Leimin Tian. 2020. Would you help a sad robot? Influence of robots' emotional expressions on human-multi-robot collaboration. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1243–1250. DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223524

[58] Joshua Zonca, Anna Folsø, and Alessandra Sciutti. 2021. The role of reciprocity in human-robot social influence. *iScience* 24, 12 (2021), 103424. DOI: https://doi.org/10.1016/j.isci.2021.103424

Appendix

A Rate of Helping Excluding Participants Who Reported Unawareness or Inability

Table A1. Percentage of Prosocial Behavior by Participants across Conditions and Rounds

		Round number			
Label	Condition	1	2	3	4
A	$R \rightarrow P, P \rightarrow H$	-	-	68.7%	76.1%
B	$H \rightarrow P, P \rightarrow H$	-	-	73.2%	75.0%
C	$H \rightarrow P, P \rightarrow R$	-	-	52.2%	65.7%
D	$R \rightarrow P, P \rightarrow R$	-	-	59.3%	66.7%
E	$P \rightarrow H$	20.8%	34.0%	43.4%	47.2%
F	$P \rightarrow R$	30.5%	33.9%	39.0%	42.4%

Empty cells indicate rounds where participants did not have an opportunity to exhibit prosocial behavior. Participants who reported unawareness or inability as motivations for not helping and never helped other agents during the experiment are excluded.

Received 26 July 2024; revised 12 April 2025; accepted 12 June 2025